PATENT ABSTRACTS OF JAPAN

(11)Publication number : 06-019962
(43)Date of publication of application : 28.01.1994

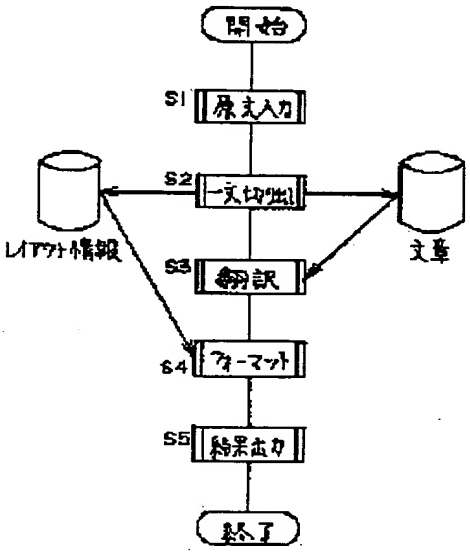
(51)Int.Cl. G06F 15/38
G06F 15/20
G06F 15/401

(21)Application number : 04-177950
(22)Date of filing : 06.07.1992

(71)Applicant : SHARP CORP
(72)Inventor : KUGIMIYA HIDEZO

(54) TEXT DIVIDING DEVICE

(57)Abstract:
PURPOSE: To accurately divide a text and to efficiently execute language processing by detecting the divided position of the text in terms of a dividing segmenting character and a format, segmenting the text on the position and outputting the segmented text.
CONSTITUTION: An original sentence to be translated is inputted S1 and one sentence segmenting processing is executed S2. The one sentence segmenting processing extracs format information such as the layout/character sorts of a text in each sentence by using segmenting characters included in the text and their format information and stores extracted layout information also correspondingly to respective sentences included in the text. The text is successively translated S3 by setting up each segmented sentence as an input unit and the formating of translated sentences is executed S4 by applying stored format information and outputted S5. Since the text is divided by using format information other than normal segmenting characters, the segmentation of a sentence which can not be expressed only by segmenting characters can be correctly segmented and the text can be segmented in each sentence.



LEGAL STATUS

[Date of request for examination] 12.07.1996
[Date of sending the examiner's decision of rejection] 06.07.1999
[Kind of final disposal of application other than the examiner's decision of rejection or application converted registration]
[Date of final disposal for application]
[Patent number]
[Date of registration]
[Number of appeal against examiner's decision of rejection]

[Date of requesting appeal against examiner's decision of rejection]

[Date of extinction of right]

(19)日本国特許庁（ＪＰ）　(12)**公開特許公報（Ａ）**　(11)特許出願公開番号

特開平6－19962

(43)公開日　平成６年(1994)１月28日

(51)Int.Cl.[5]		識別記号	庁内整理番号	ＦＩ	技術表示箇所
Ｇ０６Ｆ	15/38		Ｅ　9194－5L		
	15/20	５１４	Ａ　6798－5L		
	15/401		7218－5L		

審査請求　未請求　請求項の数１(全　６　頁)

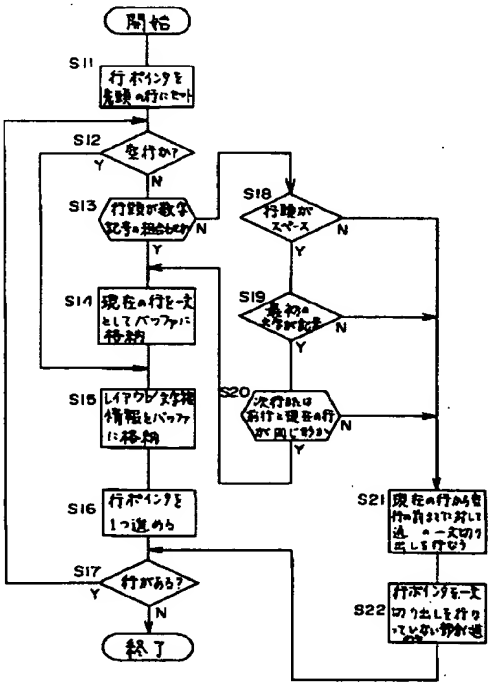
(21)出願番号　特願平4－177950

(22)出願日　平成４年(1992)７月６日

(71)出願人　000005049
シャープ株式会社
大阪府大阪市阿倍野区長池町22番22号
(72)発明者　釘宮　秀造
大阪府大阪市阿倍野区長池町22番22号　シャープ株式会社内
(74)代理人　弁理士　深見　久郎

(54)【発明の名称】　テキスト分割装置

(57)【要約】
【目的】　言語処理に先立って行なわれるテキスト分割をより正確にし、言語処理の効率を向上させる。
【構成】　テキストに含まれるピリオド、コロンなどの区切り文字と、テキストのフォーマットとから分割位置を検出するための分割位置検出部（Ｓ１１～Ｓ１３、Ｓ１８～Ｓ２０）と、検出された分割位置でテキストを区切って出力するための出力部（Ｓ１４、Ｓ２１）とを含む。



1

【特許請求の範囲】

【請求項１】　テキストに含まれる区切り文字と、前記テキストのフォーマットとから前記テキストの分割位置を検出するための分割位置検出手段と、

前記分割位置検出手段により検出された分割位置で前記テキストを区切って出力するための出力手段とを含むテキスト分割装置。

【発明の詳細な説明】

【０００１】

【産業上の利用分野】この発明は、機械翻訳、文章要約、キーワード抽出などの言語処理において用いられるテキスト分割装置に関し、特に、原文を一括入力した後、所定の分割単位、たとえば１文ごとに切出して後続する処理に出力するためのテキスト分割装置に関する。

【０００２】

【従来の技術】機械翻訳等の言語処理においては、原文テキストをＯＣＲ（光学的文字読取装置）などにより一括入力した後に所定の処理を行なうことが一般的である。この場合、機械翻訳、文章要約、キーワード抽出などの処理はテキストの１文を単位として行なわれる。そのため、一括して入力された原文テキストを１文ずつに分割する処理が必要となる。

【０００３】従来、この１文切出の処理は、ピリオド（．）、コロン（：）、セミコロン（；）などの文の切れ目を表わす区切り文字を認識することにより、この区切り文字の部分でテキストを分割して行なっていた。

【０００４】

【発明が解決しようとする課題】このような従来のテキスト分割装置では、区切り文字が存在しないとテキストをその部分で分割することができない。そのため、テキストのタイトル部分と本文部分とが分割されずひとまとめとして出力されたり、リストとして数行にわたって挙げられた多数の項目が、全体で１つの文になってしまって出力されたりするという、誤った処理が行なわれることがあった。このような誤った１文切出し処理をすると、後の処理を正しく行なうために、誤った部分を修正する作業が必要となる。そのため従来のテキスト分割装置を用いると言語処理全体の効率が悪くなるという問題点がある。

【０００５】それゆえにこの発明の目的は、従来のテキスト分割装置よりもより精度よくテキストの分割を行なうことができ、その結果後の言語処理を効率よくすることができるテキスト分割装置を提供することである。

【０００６】

【課題を解決するための手段】本発明に係るテキスト分割装置は、テキストに含まれる区切り文字と、テキストのフォーマットとからテキストの分割位置を検出するための分割位置検出手段と、分割位置検出手段により検出された分割位置で、テキストを区切って出力するための出力手段とを含む。

2

【０００７】

【作用】本発明に係るテキスト分割装置では、分割位置として区切り文字のみでなく、テキストのフォーマットをも用いて検出処理が行なわれ、このようにして検出された分割位置でテキストが分割される。

【０００８】

【発明の実施例】以下、この発明の一実施例を図面を参照して詳細に説明する。なお、本明細書においては、テキストの「フォーマット」とは、文の配列を示すレイアウトや、文を構成する各文字が用いられている文字種など、文字の配置を表わすすべての情報を示すものとする。

【０００９】図１は、本発明の一実施例に係るテキスト分割装置を用いた機械翻訳装置で行なわれる処理のフローチャートおよびハードウェアの一部を示す模式図である。まずステップＳ１で、翻訳対象の原文を図示されないＯＣＲなどにより入力する。

【００１０】続いてステップＳ２で、本発明に係るテキスト分割装置を用いて１文切出し処理が行なわれる。このときの１文切出し処理は、テキストのレイアウト／文字種などのフォーマット情報を取出し、テキストに含まれる区切り文字のみならずこれらフォーマット情報をも用いて各文ごとに行なう。このとき、抽出されたレイアウト情報も図１に示されるように文章に含まれる各文と対比させて格納する。

【００１１】ステップＳ３では、ステップＳ２で切出された１文を入力単位として順次翻訳処理を行なう。

【００１２】続いてステップＳ４で、ステップＳ３で得られた翻訳結果の文に対して、ステップＳ２の処理によって一旦格納していたレイアウト／文字種などのフォーマット情報を適用して訳文のフォーマッティングを行なう。

【００１３】そしてステップＳ５で、ステップＳ４でフォーマットされた結果の文を出力して終了する。

【００１４】図２は、図１のステップＳ２で行なわれる１文切出し処理のより詳細な手順を示すフローチャートである。図３は入力テキストの一例を示す模式図であり、図４は図３に対して１文切出し処理を行なった場合に得られたレイアウト／文字種情報と、切出された文との対応関係を示すバッファの模式図である。

【００１５】図２を参照して、１文切出しは次のようにして行なわれる。まずステップＳ１１で、テキストのうちの処理対象となっている行を指す行ポインタを、テキストの先頭行にセットする処理が行なわれる。

【００１６】ステップＳ１２で、ポインタの指す行が空行かどうかを判定する処理が行なわれる。空行とは、何も文字が含まれていない行を指す。処理対象の行が空行である場合には処理はステップＳ１５に進み、空行でない場合には処理はステップＳ１３に進む。

【００１７】ステップＳ１３では、処理対象の行の先頭

が数字と記号の組合せであるかどうかについての判断が行なわれる。行頭が数字と記号の組合せである場合にはこの行はタイトルである可能性が高い。そのため処理はステップＳ１４に進む。行頭が数字と記号の組合せでない場合には処理はステップＳ１８に進む。

【００１８】ステップＳ１４では、現在処理中の行を１文（１つの単位）としてバッファに格納する処理が行なわれる。ステップＳ１４の後処理はステップＳ１５に進む。

【００１９】ステップＳ１３からステップＳ１８に処理が進んだ場合、ステップＳ１８では、処理対象の行の行頭がスペースであるかどうかについての判定が行なわれる。行頭がスペースであれば処理はステップＳ１９に進み、それ以外の場合には処理はステップＳ２１に進む。

【００２０】ステップＳ１９では、スペースの後の最初の文字が記号であるかどうかについての判断が行なわれる。記号である場合には処理はステップＳ２０に、それ以外の場合には処理はステップＳ２１に進む。

【００２１】ステップＳ２０においては、現在の行の次の行または現在の行の前の行と現在の行とが同じ形かどうかについての判断が行なわれる。同じ形かどうかとは、行頭がスペースであってかつ最初の文字が記号であるか、あるいはそうした条件が成立しないかということである。次行または前行が現在の行と同じ形の場合には処理はステップＳ１４に進み、それ以外の場合には処理はステップＳ２１に進む。ステップＳ１４に処理が進んだ場合、行頭が数字と記号の組合せであった場合と同様に現在の行を１文としてバッファに格納する処理が行なわれ、さらにステップＳ１５以下に処理が進む。

【００２２】一方ステップＳ１８、ステップＳ１９、ステップＳ２０の３つの判断のいずれかでＮＯという判断が行なわれた場合処理はステップＳ２１に進む。ステップＳ２１では、現在の行から、次の空行の前までに対して、通常の１文切出し処理を施す。すなわち、テキストに含まれるピリオドやコロンなどの区切り文字でテキストを分割し、それぞれを１文として処理を行なう。処理はステップＳ２２に進む。

【００２３】ステップＳ２２では、行ポインタを、まだ１文切出し処理を行なっていない部分まで進める処理を行なう。ステップＳ２２の後処理はステップＳ１７に進む。

【００２４】一方、ステップＳ１５では、レイアウト／文字種などのテキストのフォーマット情報をバッファに格納する処理が行なわれる。ここでフォーマット情報としては、文頭にスペースがある場合のそのスペースの個数、使用されている活字の種類（たとえばボールド体、イタリック体など）、文末に改行があるどうかなどの情報を含む。この詳細については図３、４を参照して後に説明する。

【００２５】ステップＳ１５の後処理はステップＳ１６に進み、行ポインタを１つ進める処理が行なわれる。これにより処理対象の行は１つ先に進むことになる。ステップＳ１６の後処理はステップＳ１７に進む。

【００２６】ステップＳ１７では、ステップＳ１６、ステップＳ２２で新たに設定された行ポインタで示される位置に、処理対象となる行が存在するかどうかについての判断が行なわれる。存在する場合には処理は再びステップＳ１２に戻りステップＳ１２以下の処理が繰返して実行される。行が存在しない場合には処理は終了する。

【００２７】図２に示されるような１文切出し処理を行なうことにより、次のような結果を得ることができる。図３は、入力テキストの一例である。図３に示されるテキストの場合には、タイトルと、本文とが空白行で分離されている。また本文はさらに、地の文を表わす部分と、この地の文によって導入される多数の例示部分とが含まれ、これら２つの部分は空行で分離されている。

【００２８】図３に示されるテキストの場合には、通常の区切り文字以外の部分でテキストを分割しなければ、たとえばタイトルと地の文の部分が相互に接続されてしまったり、例示の文が相互に複数個接続されてしまったりし、正しい１文切出し処理が行なわれない。

【００２９】これに対し、本願発明のテキスト分割装置を用いてこの文を分割すると、その結果は図４に示されるようになる。図４を参照して、文ナンバー１のタイトルと文ナンバー３の地の文とは、文ナンバー２の空行によって分離されている。また文ナンバー３と文ナンバー４とは通常の区切り文字（ピリオド）により分離され、文ナンバー４と文ナンバー６との間は通常の区切り文字（コロン）および文ナンバー５の空行によって分離されている。また文ナンバー６、７の例示の部分は、文末に改行が存在することからこのレイアウト情報によって２つの文に分割される。他の例示の文も同様に分離される。また文ナンバー１と文ナンバー３との間では、使用されている文字種が異なっていることを用いても分割が可能である。

【００３０】以上のように本発明に係るテキスト分割装置では、通常の区切り文字以外のフォーマット情報を用いてテキストの分割が行なわれる。そのための、区切り文字のみでは表わせないような文の区切りを正しく検出してテキストを１文ずつに切出す処理が可能である。区切り文字のみでは分割不能な文も正しく分割することができるため、後続する言語処理に先立って１文切出し処理の結果を修正する必要性は少なく、処理の効率を向上させることができる。

【００３１】

【発明の効果】以上のように本発明に係るテキスト分割装置には、通常の区切り文字のみでは表現できないテキストの分割位置を、テキストのフォーマット情報を用いて検出し、このように検出された分割位置でテキストを分割することができる。そのため、区切り文字のみを用

5

いてテキスト分割を行なった場合に比べてテキスト分割の精度がより向上し、後続する処理に先立ってテキスト分割の処理結果を訂正する必要性は少なくなる。

【００３２】その結果、テキスト分割の精度をより向上させることができ、かつ後続する言語処理の効率も高めることができるテキスト分割装置を提供できる。

【図面の簡単な説明】

【図１】本発明の一実施例に係るテキスト分割装置を用＊

6

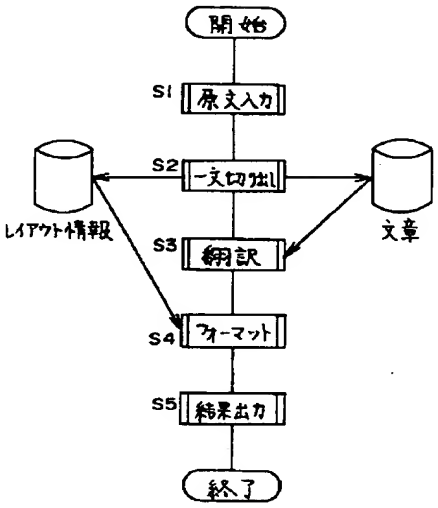
＊いた機械翻訳装置で行なう処理のフローチャートおよび装置の一部を示す模式図である。

【図２】１文切出処理のフローチャートである。

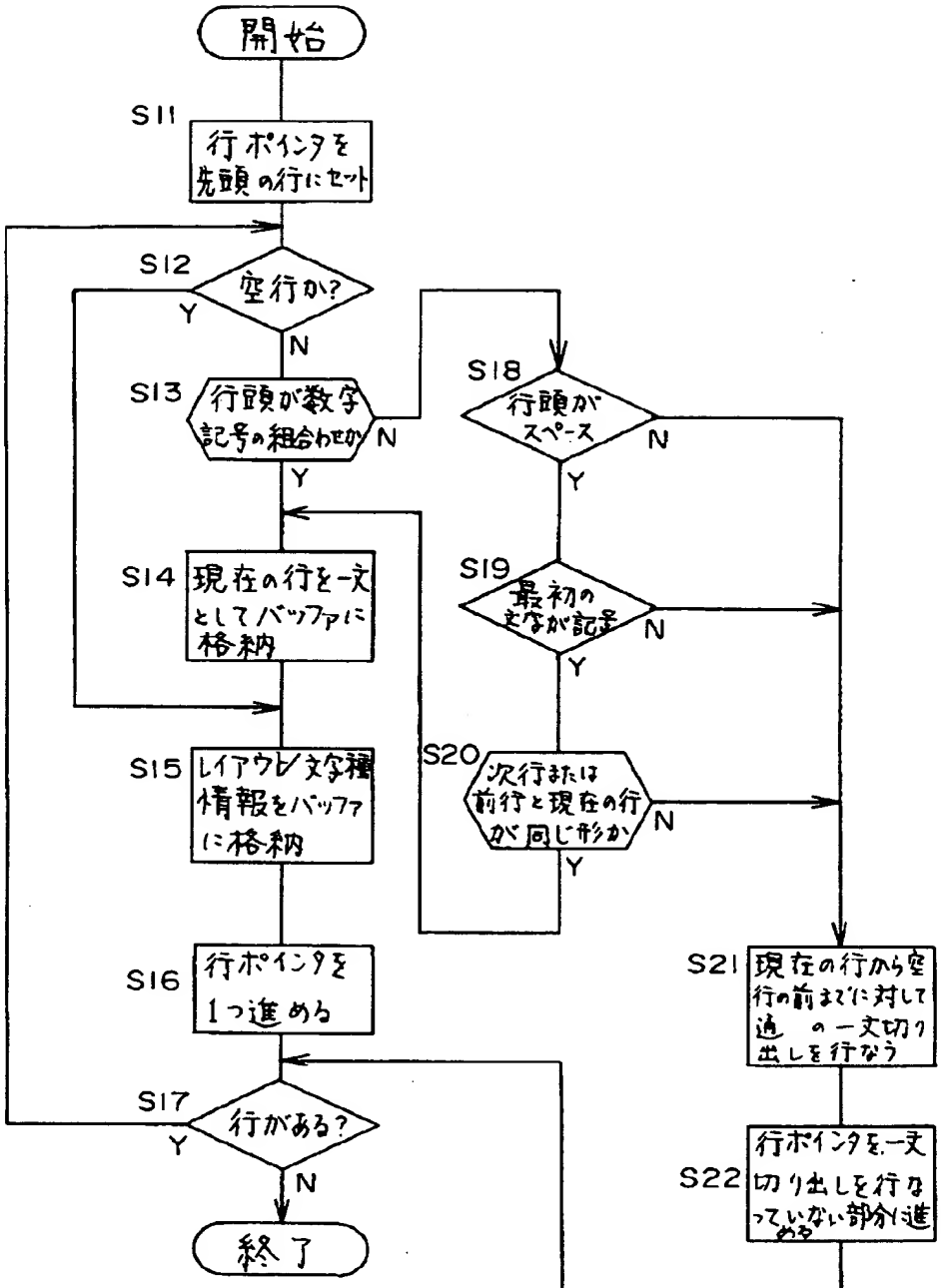
【図３】入力テキストの一例を示す模式図である。

【図４】図３に示されるテキストを本発明に係るテキスト分割装置で分割した場合の処理結果を示すバッファの模式図である。

【図１】



【図２】


開始
S11
行ポインタを先頭の行にセット
S12
空行か?
Y
N
S13
行頭が数字記号の組合わせか
N
Y
S14
現在の行を一文としてバッファに格納
S15
レイアウト文字種情報をバッファに格納
S16
行ポインタを1つ進める
S17
行がある?
Y
N
終了
S18
行頭がスペース
N
Y
S19
最初の文字が記号
N
Y
S20
次行または前行と現在の行が同じ形か
N
Y
S21
現在の行から空行の前までに対して一文切り出しを行なう
S22
行ポインタを一文切り出しを行なっていない部分に進める

5.4.1.3 SPECIALIZATION OF PERSONNEL

A machine translation chain necessitates at the outset the following basic personnel: computer scientist, linguist, translator and terminologist. Each of these persons should have a specialty: for example:

- a computer scientist specializing in natural language processing
- a computer scientist specializing in data bases
- a computer scientist specializing in software engineering
- a computer scientist specializing in telecommunications
- a linguist specializing in computational linguistics
- a linguist specializing in theoretical linguistics
- a linguist specializing in contrastive linguistics
- a translator specializing in the particular domain
- a translator specializing in machine translation
- a translator specializing in lexicography
- a translator specializing in terminology.

【図3】

文 No.	レイアウト／文字種情報	文
1	ＢＯＬＤ体／文末で改行	5.4.1.3 SPECIALIZATION OF PERSONNEL
2	空行	
3	文頭スペース４個	A machine translation chain necessitates at the outset the following basic personell: computer scientists, linguist, translator and terminologist.
4	文末で改行	Each of these persons should have a specialty: for example:
5	空行	
6	文頭にスペース４個 文末で改行	- a computer scientist specializing in natural language processing
7	文頭にスペース４個 文末で改行	- a computer scientist specializing in data bases

【図4】

特開平6－19962